Assignment 5 Report for Part A

Travis Xie

0a. Bellman Equations:

$$Q^*(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

$$V^*(s) = \max_a Q^*(s,a)$$

0b. Bellman Update:

$$Q_{k+1}(s,a) \leftarrow \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$$

$$V_{k+1}(s) \leftarrow \max_a Q_k(s,a)$$

1a. How many iterations of VI are required to turn 1/3 of the states green? (i.e., get their expected utility values to 100). 4 Iterations are required.

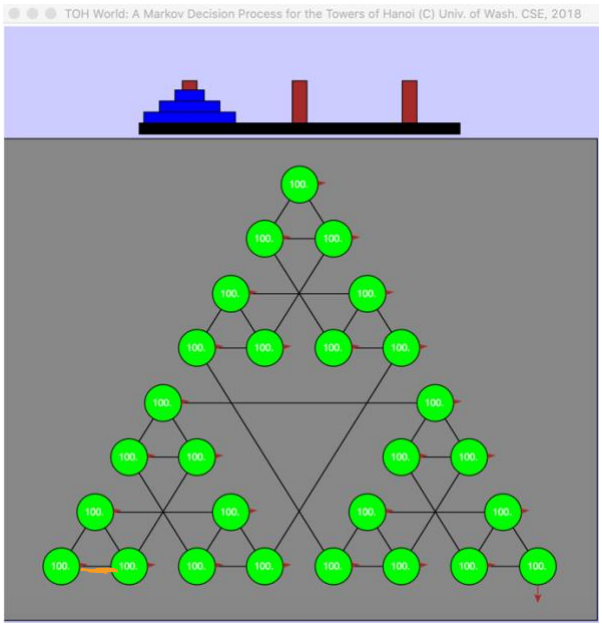
1b. How many iterations of VI are required to get all the states, including the start state, to 100?

8 iterations.

1c. From the Value Iteration menu, select "Show Policy from VI". (The policy at each state is indicated by the outgoing red arrowhead. If the suggested action is illegal, there could still be a legal state transition due to noise, but the action could also result in no change of state.) Describe this policy. Is it a good policy? Explain.

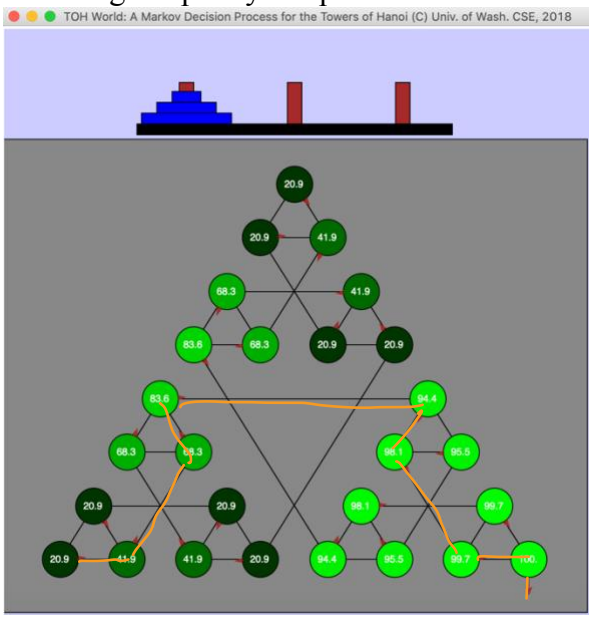
Policy: Move disk 1 to peg3, Illegal action (no change of state)

This is not a good policy because it never leads the agent to the goal.



2a. How many iterations are required for the start state to receive a nonzero value. 8 Iterations

2c. At this point, view the policy from VI as before. Is it a good policy? Explain.



Yes, it is a good policy. Actually, it is the optimal policy because it takes the least steps to reach

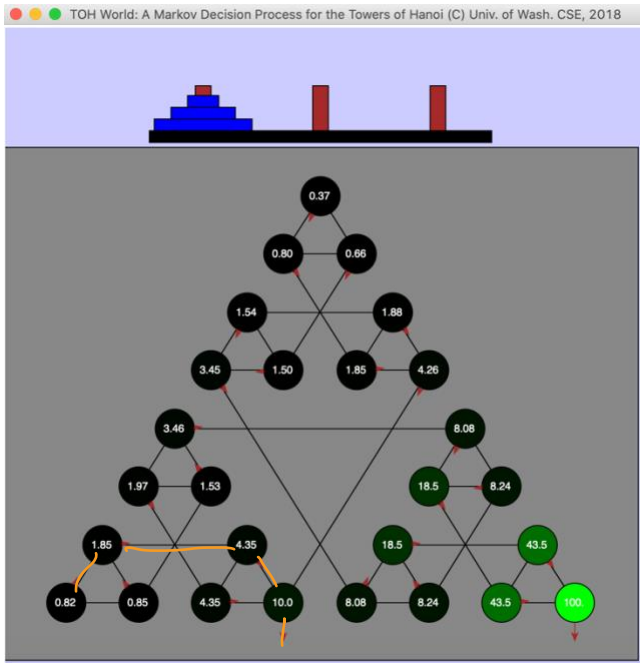
the final state.

2d. Run additional VI steps to find out how many iterations are required for VI to converge. How many is it? 56 Iterations

2e. After convergence, examine the computed best policy once again. Has it changed? If so, how? If not, why not? Explain.

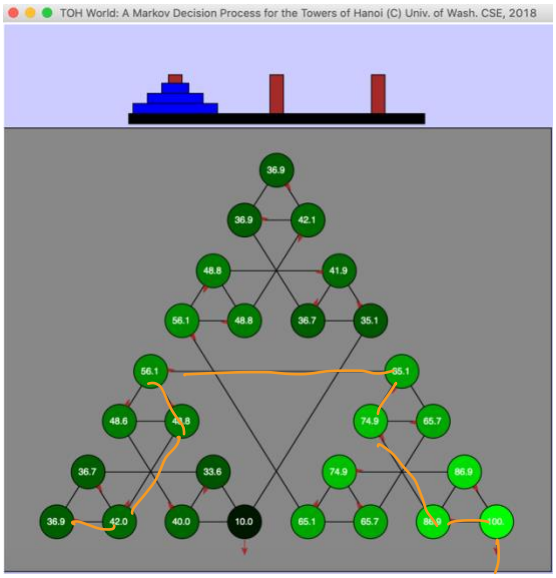
No, the best policy has not changed because the best policy has already appeared after 8 iterations.

3a. Run Value Iteration until convergence. What does the policy indicate? What value does the start state have? (start state value should be 0.82)



This policy always leads the game to reach the goal of R = 10. The start state has the value of 0.82

3b. Reset the values to 0, change the discount to 0.9 and rerun Value Iteration until convergence. What does the policy indicate now? What value does the start state have? (start state value should be 36.9)



Now, the policy will lead us to reach the goal of R = 100, which is the goal we want. The start state has the value of 36.9.

4a. In how many of these simulation runs did the agent ever go off the plan?

The agent goes off the plan in every simulation since there is a noise.

4b. In how many of these simulation runs did the agent arrive in the goal state (at the end of the golden path)? Only two out of ten simulations arrived in the goal state.

4c. For each run in which the agent did not make it to the goal in 10 steps, how many steps away from the goal was it? 2, 3, 1, 3, 4, 2, 1, 2. On average, 2.25 steps away from the goal.

4d. Are there parts of the state space that seemed never to be visited by the agent? If so, where (roughly)? Yes, it seems like that the agent never explored the top part of the state space.

5a. Since it is having a good policy that is most important to the agent, is it essential that the values of the states have converged?

No, it is not essential because the best policy might appear way before convergence.

5b. If the agent were to have to learn the values of states by exploring the space, rather than computing with the Value Iteration algorithm, and if getting accurate values requires re-visiting states a lot, how important would it be that all states be visited a lot? Explain

We want to explore new states and revisit states a lot, but it is not too important to visit all states because that all states in state space are related to the optimal policy. For example, states in the top part of state space of TOH are rarely visited by the agent so it will be a waste of time to visit them in learning.